# Robust Learning from Bites for Data Mining ⋆

Andreas Christmann *

*Vrije Universiteit Brussel, Department of Mathematics, BELGIUM*

Ingo Steinwart

*Los Alamos National Laboratory, Los Alamos, NM 87545, CCS-3, USA*

Mia Hubert

*Katholieke Universiteit Leuven, Department of Mathematics, BELGIUM*

**Abstract**

Some methods from statistical machine learning and from robust statistics have two drawbacks. Firstly, they are computer-intensive such that they can hardly be used for massive data sets, say with millions of data points. Secondly, robust and non-parametric confidence intervals for the predictions according to the fitted models are often unknown. A simple but general method is proposed to overcome these problems in the context of huge data sets. An implementation of the method is scalable to the memory of the computer and can be distributed on several processors to reduce the computation time. The method offers distribution-free confidence intervals for the median of the predictions. The main focus is on general support vector machines (SVM) based on minimizing regularized risks. As an example, a combination of two methods from modern statistical machine learning, i.e. kernel logistic regression and $\varepsilon$-support vector regression, is used to model a data set from several insurance companies. The approach can also be helpful to fit robust estimators in parametric models for huge data sets.

*Key words:* Breakdown point, convex risk minimization, data mining, distributed computing, influence function, logistic regression, robustness, scalability, statistical machine learning, support vector machine.

⋆ We thank Peter Rousseeuw, Claudio Agostinello, Matias Salibian-Barrera, and Stefan Van Aelst for helpful discussions during ICORS-2005 in Jyväskylä.

* Corresponding author: Andreas Christmann, Vrije Universiteit Brussel (VUB), Department of Mathematics, Pleinlaan 2, B–1050 Brussel, BELGIUM.

*Email address:* Andreas.Christmann@vub.ac.be (Andreas Christmann).

*URL:* homepages.vub.ac.be/∼achristm/ (Andreas Christmann).

## 1 Introduction

Data sets with millions of observations occur nowadays in many areas, *e.g.* insurance companies or banks collect many variables to develop tariffs and scoring methods for credit risk management, respectively. Other examples are large observational data sets in data mining projects and data from micro-arrays. Although such big data sets contain a lot of valuable information, the analysis of such data sets can be cumbersome due to limited computer memory or computational time problems. Classical parametric assumptions are often violated for such data sets which probably contain some outliers. We give only three citations for these facts. Hampel *et al.* (1986, p. 27f) made the following comment on data quality and gross errors. *"There are often no or virtually no gross errors in high-quality data, but* 1% *to* 10% *of gross errors in routine data seem to be more the rule than the exception"*. J.W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 (cited from Hampel *et al.* (1986, p. 21)): *"A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians."* Le Cam (1980, p.478) concluded for data sets with $n = 10^5$ to $n = 10^8$ data points: *"Thus the asymptotics fail precisely when one would feel that they are applicable."* Hence, it is no surprise that the data quality in large data mining projects is often far from being optimal, *cf.* Hand *et al.* (2001) or Hipp *et al.* (2001), and the application of robust statistical methods is therefore important in such situations.

Unfortunately, many robust methods proposed in the literature have the following drawbacks which are serious limitations for their application. (a) They are computer-intensive such that they can hardly be used for massive data sets, say for several millions of observations with hundreds of explanatory variables. (b) Robust standard errors and robust confidence intervals for the estimated parameters or for robust predictions are often unknown. (c) Some statistical software packages like S-PLUS or R contain state-of-the-art algorithms for robust statistical methods, but the implemented numerical algorithms usually require that the whole data set fits into the memory of the computer.

In this paper a simple but quite general method for robust estimation in the context of huge data sets is proposed. The main goal of the proposal is to broaden the application of robust general SVM methods for massive data. The idea is to partition the huge data set $S$ at random into disjoint subsets $\mathcal{S}_b$, $b = 1, \ldots, B$. Then a robust method is applied to each subset, and the results are summarized in a robust manner. The proposal yields robust predictions. If the median is used to aggregate the $B$ single predictions then we also get

robust and distribution-free confidence intervals.

The rest of the paper is organized as follows. Section 2 gives the proposed method and Section 3 describes its properties. Section 4 gives some numerical examples for the case of robust linear regression and kernel logistic regression. Section 5 contains a summary and compares RLB with competing methods. All proofs are given in the Appendix.

## 2 Method

In this section we describe a simple but rather general method for robust estimation for huge data sets. We restrict attention to classification and regression problems although the method can be used in other fields as well. The proposal has two goals: making robust general SVM methods usable for data sets which are too large for currently available algorithms due to memory or time limitations and offering robust and distribution-free confidence intervals based on the median for the predictions.

In classification and in regression problems one assumes an approximate functional relationship between an explanatory random variable $X$ and a response random variable $Y$ using $n$ observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ drawn independently from the same probability distribution P of the pair $(X, Y)$. In a non-parametric setting the distribution P is totally unknown. For technical reasons we assume throughout this work that $\mathcal{X}$ and $\mathcal{Y}$ are closed or open subsets of $\mathbb{R}^d$ and $\mathbb{R}$, respectively. Hence we can split up P into the marginal distribution $\mathrm{P}_X$ and the regular conditional probability $\mathrm{P}(\cdot\,|x)$, $x \in \mathcal{X}$, on $\mathcal{Y}$. For the case of binary classification we have $\mathcal{Y} = \{-1, +1\}$.

Under the classical signal plus noise assumption $Y_i|(X = x_i)$ is distributed as $f(x_i) + \varepsilon_i$, where $f$ is an unknown function and $\varepsilon_i$ are independent and identically distributed error terms, $1 \leq i \leq n$. In the parametric setup we have $f(x) = f_\theta(x) = x'\theta$, $\theta \in \Theta \subset \mathbb{R}^d$. In the non-parametric setup $f$ belongs to some Hilbert space $\mathcal{H}$ of all measurable functions $f : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In our case $\mathcal{H}$ will be a (typically infinite dimensional) reproducing kernel Hilbert space (RKHS).

In this paper we always assume that the sample size $n$ is large. The whole data set is often partitioned at random into two or three disjoint parts for training, validation, and testing purposes. Instead of modeling the full training data set, we split the training data set at random into $B \geq 1$ parts $\mathcal{S}_b$ (called 'bites') of approximately the same sub-sample sizes $n_b \approx n/B$. Then we fit each bite with the robust method. Finally, we compute a robust location estimator of the estimators $T_{\mathcal{S}_b}$ and summarize the predictions from the $B$ fitted models.

**Definition 1** *Let $\mathcal{S} = ((x_1, y_1), \ldots, (x_n, y_n))$ be a sample of size $n$ from a probability distribution* P *on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$. Let $T_{\mathcal{S}}$ be the $\mathcal{H}$-valued estimator of interest. Consider a random partition of $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_B$ into $B$ non-empty disjoint subsets, where $n_b := |\mathcal{S}_b| \in \mathbb{N}$, $n = \sum_{b=1}^{B} n_b$, $b = 1, \ldots, B$, $B \in \{1, \ldots, n\}$, $B \ll n$. An **RLB estimator of type I** is defined by*

$$T_{\mathcal{S},B}^{RLB} = g(T_{\mathcal{S}_1}, \ldots, T_{\mathcal{S}_B}), \tag{1}$$

*where $g : \mathcal{H}^B \to \mathcal{H}$ is a measurable map. An **RLB estimator of type II** is given by*

$$T_{\mathcal{S},B}^{RLB}(x) = g^*(T_{\mathcal{S}_1}(x), \ldots, T_{\mathcal{S}_B}(x)), \qquad \forall\, x \in \mathcal{X}, \tag{2}$$

*where $g^* : \mathbb{R}^B \to \mathbb{R}$ is a measurable map.*

**Remarks.** (i) An RLB estimator of type I can obviously be used to define an RLB estimator of type II. (ii) An RLB estimator of type II does not necessarily define an RLB estimator of type I, because the related function $g^*$ does not necessarily correspond to a function $g$ mapping onto the Hilbert space $\mathcal{H}$. (iii) The class of RLB estimators of type I − and due to part (i) of this remark the class of RLB estimators of type II − is non-empty, because we obtain $g(T_{\mathcal{S}_1}, \ldots, T_{\mathcal{S}_B}) := \frac{1}{B} \sum_{b=1}^{B} T_{\mathcal{S}_b} \in \mathcal{H}$ for $g$ equal to the mean. ◁

We will often consider RLB estimators of type I with are *convex* combinations

$$T_{\mathcal{S},B}^{RLB} = \sum_{b=1}^{B} c_b T_{\mathcal{S}_b} \tag{3}$$

with weights $c_b \in (0, 1)$ and $\sum_{b=1}^{B} c_b = 1$ ($c_b \equiv \frac{1}{B}$ gives the mean), and RLB estimators of type II based on the median. The convexity assumption will assure that the RLB estimator belongs to the set of valid solutions provided the parameter space is a convex set. This is true e.g. if the parameter space is equal to $\mathbb{R}^d$ in a parametric situation or if the parameter space is a Hilbert space $\mathcal{H}$ for kernel based methods. Of course, L–estimators such as $\alpha$-trimmed means, M–, S–, and R–estimators can also be used in the aggregation step.

If $B$ is large enough, say above 15, precision estimates can additionally be obtained by computing standard deviations of the predictions $T_{\mathcal{S},B}^{RLB}(x)$ using the central limit theorem. However, in general we favor a distribution-free method based on the median. If $B$ is small or if the distribution or the variance of $T_{\mathcal{S},B}^{RLB}(x)$ is unknown, one can construct distribution-free confidence intervals for the median of $T_{\mathcal{S},B}^{RLB}(x)$ and distribution-free tolerance regions based on selected order statistics, see David and Nagaraja (2003, Chap. 7). Table 1 lists some values of $B$, the corresponding pair of order statistics determining the confidence interval, the lower bound of the actual confidence level which is $0.5^B \sum_{j=r}^{s} \binom{B}{j}$, and the finite sample breakdown point (see Definition 12) $\varepsilon_B^* = \min\{r - 1, B - s\}/B$ of the confidence interval. In Section 3 it will be shown that RLB inherits robustness properties from the original estimator

and from the estimator used in the aggregation step. The actual confidence intervals based on the median can be conservative of small choices of $B$, see Table 1. If $B$ is not too small, say $B > 15$, this breakdown point is high enough for most practical applications. E.g. fix $B = 17$. Then the $5^{th}$ and the $13^{th}$ order statistics give a confidence interval at the level 95% for the median which is valid for *all* distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The breakdown point of this confidence interval is $4/17 = 0.235$ because the values of the four lowest and the four highest predictions are not used.

| $1-\alpha$ | $B$ | $r$ | $s$ | lower bound of confidence level | finite sample breakdown point |
|---|---|---|---|---|---|
| 0.90 | 8 | 2 | 7 | 0.930 | 0.125 |
| | 10 | 2 | 9 | 0.979 | 0.100 |
| | 13 | 4 | 10 | 0.908 | 0.231 |
| | 18 | 6 | 13 | 0.904 | 0.278 |
| | 30 | 11 | 20 | 0.901 | 0.333 |
| | 53 | 21 | 33 | 0.902 | 0.377 |
| | 104 | 44 | 61 | 0.905 | 0.413 |
| 0.95 | 9 | 2 | 8 | 0.961 | 0.111 |
| | 10 | 2 | 9 | 0.979 | 0.100 |
| | 17 | 5 | 13 | 0.951 | 0.235 |
| | 37 | 13 | 25 | 0.953 | 0.324 |
| | 51 | 19 | 33 | 0.951 | 0.353 |
| | 58 | 22 | 37 | 0.952 | 0.362 |
| | 101 | 41 | 61 | 0.954 | 0.396 |
| 0.99 | 10 | 1 | 10 | 0.998 | 0.000 |
| | 12 | 2 | 11 | 0.994 | 0.083 |
| | 26 | 7 | 20 | 0.991 | 0.231 |
| | 39 | 12 | 28 | 0.991 | 0.282 |
| | 49 | 16 | 34 | 0.991 | 0.306 |
| | 101 | 38 | 64 | 0.991 | 0.366 |

Table 1
Selected pairs $(r, s)$ of order statistics for non-parametric confidence intervals at the $(1-\alpha)$-level for the median.

## 3 Properties of RLB

In this section properties of robust learning from bites are investigated. Denote the estimator based on the whole data set by $T_{\mathcal{S}}$ and denote the corresponding RLB estimator based on $B$ bites by $T_{\mathcal{S},B}^{RLB}$. We will assume in this section that $\min_{1\le b\le B} n_b \to \infty$, $n \to \infty$. Computational time and memory space are considered in Section 3.1. RLB for general SVM estimators is investigated in

Section 3.2, and robustness properties are proved in Section 3.3. In Section 3.4 some arguments are given how to choose the number of bites. All proofs are given in the appendix.

*3.1 General properties*

The estimators $T_{\mathcal{S}_b}$, $1 \leq b \leq B$, from the bites are stochastically independent because $\mathcal{S}_1, \ldots, \mathcal{S}_B$ are disjoint. Denote the number of available CPUs by $k$ and let $k_B$ be the smallest integer which is not smaller than $B/k$.

**Proposition 2 ($k$ CPUs)** *Assume that the computation time of $T_{\mathcal{S}}$ for a data set with $n$ observations and $d$ explanatory variables is of order $O(h(n,d))$, where $h$ is some positive function. Then the computation time of $T_{\mathcal{S},B}^{RLB}$ with subsample sizes $n_b \approx n/B$ is approximately of order $O(k_B \cdot h(n/B, d))$.*

**Proposition 3 ($k$ CPUs)** *Assume that the estimator $T_{\mathcal{S}}$ for a data set with $n$ observations and $d$ explanatory variables needs memory space and hard disk space of order $O(h_1(n,d))$ and $O(h_2(n,d))$, respectively, where $h_1$ and $h_2$ are positive functions. Then the computation of $T_{\mathcal{S},B}^{RLB}$ for subsample sizes $n_b \approx n/B$ needs approximately memory space and hard disk space of order $O(k \cdot h_1(n/B, d))$ and $O(k \cdot h_2(n/B, d))$, respectively.*

**Proposition 4 (Consistency)** *Consider an RLB estimator $T_{\mathcal{S},B}^{RLB}$ of type I based on a convex combination with $c_b \in (0,1)$ and $\sum_{b=1}^{B} c_b = 1$. (i) If $\mathbb{E}(T_{\mathcal{S}_b}) = \mathbb{E}(T_{\mathcal{S}})$ for all $b \in \{1, \ldots, B\}$, then $\mathbb{E}(T_{\mathcal{S},B}^{RLB}) = \mathbb{E}(T_{\mathcal{S}})$. (ii) If $T_{\mathcal{S}}$ converges in probability (or almost sure) to $T_{\mathrm{P}}$ for $n \to \infty$ and if $(n/n_b) \to B$, $B$ fixed, then $T_{\mathcal{S},B}^{RLB}$ converges in probability (or almost sure) to $T_{\mathrm{P}}$. (iii) Let $c_b \equiv \frac{1}{B}$. Assume that $n_b^{1/2}(T_{\mathcal{S}_b} - T_{\mathrm{P}})$ converges weakly to a multivariate normal distribution $N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, and that $(n/n_b) \to B$, $1 \leq b \leq B$, $B$ fixed. Then $n^{1/2}(T_{\mathcal{S},B}^{RLB} - T_{\mathrm{P}})$ converges weakly to a multivariate normal distribution $N(0, \Sigma)$, $n \to \infty$.*

**Proposition 5 (Consistency)** *Consider an RLB estimator $T_{\mathcal{S},B}^{RLB}$ of type II where the **median** is used in the aggregation step. If $T_{\mathcal{S}}(x)$ converges in probability (or almost sure) to $T_{\mathrm{P}}(x)$, $x \in \mathcal{X}$, and if $\lim_{n\to\infty} (n/n_b) \equiv B$, $B$ fixed, then $T_{\mathcal{S},B}^{RLB}(x)$ converges in probability (or almost sure) to $T_{\mathrm{P}}(x)$, $x \in \mathcal{X}$.*

*3.2 Properties of RLB using the mean for general SVM methods*

Now we consider general SVM estimators

$$f_{\mathcal{S},\lambda} := \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda||f||^2_{\mathcal{H}}\,, \tag{4}$$

for $f$ where $L : \mathcal{Y} \times [0,\infty) \to [0,\infty)$ is a convex loss function, $\mathcal{H}$ is the reproducing kernel Hilbert space defined via the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and regularizing parameter $\lambda > 0$, see Vapnik (1998) and Schölkopf and Smola (2002). Special cases of such general SVM methods are the *support vector machine*: $L(y,t) = \max\{0, 1-yt\}$, $y \in \{-1,+1\}$, $t \in \mathbb{R}$, *kernel logistic regression*: $L(y,t) = \ln(1+\exp[-yt])$, $y \in \{-1,+1\}$, $t \in \mathbb{R}$, and *support vector regression*: $L(y,t) = \max\{|y-t| - \varepsilon, 0\}$, $y,t \in \mathbb{R}$, where $\varepsilon > 0$ is fixed. The general SVM estimator $f_{\mathcal{S}_b,\lambda}(x)$, $x \in \mathcal{X}$, defined as the solution of (4) for bite $\mathcal{S}_b$ is a kernel based estimator and can be written as

$$f_{\mathcal{S}_b,\lambda}(x) = \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i), \quad i \in \mathcal{S}_b,\ x \in \mathcal{X}, \tag{5}$$

where $\alpha_{i,b} \in \mathbb{R}$. If $\alpha_{i,b} \neq 0$, then $(x_i, y_i)$ is called a support vector (SV). We denote the set of support vectors in bite $\mathcal{S}_b$ by $SV(\mathcal{S}_b)$ and its number of elements by $|SV(\mathcal{S}_b)|$. Obviously, the minimization problem (4) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk

$$f_{\mathrm{P},\lambda} := \arg\min_{f\in\mathcal{H}} \mathrm{E_P}\, L(Y, f(X)) + \lambda\|f\|^2_{\mathcal{H}} \in \mathcal{H}\,. \tag{6}$$

**Theorem 6 (RLB for general SVMs)** *Assume that the estimator $f_{\mathcal{S},\lambda}$ is a general SVM estimator defined by (4) for the whole data set with $n = \sum_{b=1}^{B} n_b$ observations, $B$ fixed. Consider an RLB estimator of type I based on a convex combination with $c_b \in (0,1)$ and $\sum_{b=1}^{B} c_b = 1$. Then the RLB estimator is itself a kernel based estimator and can be written as*

$$f^{RLB}_{\mathcal{S},B,\lambda}(x) = \sum_{i=1}^{n} \alpha_{i,RLB}\, k(x, x_i) \tag{7}$$

$$= \sum_{i\,\in\, SV(\mathcal{S}_1)\,\cup\,\ldots\,\cup\, SV(\mathcal{S}_B)} \alpha_{i,RLB}\, k(x, x_i), \quad x \in \mathcal{X}, \tag{8}$$

*where $\alpha_{i,RLB} = \sum_{b=1}^{B} c_b\, \alpha_{i,b}$ , $i = 1, \ldots, n$.*

If all support vectors are different, we have $\alpha_{i,RLB} = c_b\, \alpha_{i,b}$ in (8).

Let us now investigate the *number of support vectors* in more detail for the case of pattern recognition, *i.e.* $\mathcal{Y} = \{-1,+1\}$. For part (ii) of Theorem 10

we need some notations. Let $\mathcal{X}_0 := \{x \in \mathcal{X}; \mathrm{P}(Y = 1|X = x) = 1/2\}$ and $\mathcal{X}_{cont} := \{x \in \mathcal{X}; \mathrm{P}_X(\{x\} = 0)\}$.

**Definition 7** *Let $\mathcal{H}$ be a Hilbert space, $F : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in \mathcal{H}$ with $F(w) \neq \infty$. Then the* ***subdifferential*** *of $F$ at $w$ is defined by*

$$\partial F(w) := \Big\{ w^* \in \mathcal{H} : \langle w^*, v - w \rangle \leq F(v) - F(w) \quad \textit{for all } v \in \mathcal{H} \Big\}.$$

*Furthermore, if $L$ is a convex loss function, we denote the subdifferential of $L$ with respect to the second variable by $\partial_2 L$.*

Further, define the set-valued function

$$F_L^*(\alpha) := \Big\{ t \in \overline{\mathbb{R}}; [\alpha L(1,t) + (1-\alpha) L(-1,t)] = \min_{s \in \overline{\mathbb{R}}} [\alpha L(1,s) + (1-\alpha) L(-1,s)] \Big\},$$

$\alpha \in [0,1]$, $\partial f(A) := \cup_{a \in A} \partial f(a)$, the set

$$S = \Big\{ (x,y) \in \mathcal{X}_{cont} \times \mathcal{Y}; 0 \notin \partial_2 L(y, F_L^*(\mathrm{P}(Y = 1|X = x))) \cap \mathbb{R} \Big\},$$

and the quantity

$$S_{L,\mathrm{P}} = \begin{cases} \mathrm{P}(S) & \text{if } 0 \notin \partial_2 L(1, F_L^*(0.5)) \cap \partial_2 L(-1, F_L^*(0.5)) \\ \mathrm{P}(S) + \frac{1}{2} \mathrm{P}_{\mathcal{X}}(\mathcal{X}_0 \cap \mathcal{X}_{cont}) & \text{else,} \end{cases}$$

see Steinwart (2003, p.1082). A loss function is called *classification calibrated* if for every $\alpha \in [0,1]$ we have $F_L^*(\alpha) \subset [-\infty, 0)$ if $\alpha < 1/2$, and $F_L^*(\alpha) \subset (0, \infty]$ if $\alpha > 1/2$. Such loss functions were called admissible by Steinwart (2003), but we think that the notion of classification calibrated is more precise. For more information on this and related concepts we refer to Steinwart (2005b) and Bartlett *et al.* (2006). We also need the notion of a universal kernel proposed by Steinwart (2001) to describe the richness of the RKHS $\mathcal{H}$. We refer to Steinwart *et al.* (2006) for some more general notions and related results.

**Definition 8** *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel with reproducing kernel Hilbert space $\mathcal{H}$. Then $k$ is* ***universal*** *if $\mathcal{H}$ is dense in the space of continuous functions $C(\mathcal{X})$ equipped with $\| \, . \, \|_\infty$, i.e. for every continuous function $g : \mathcal{X} \to \mathbb{R}$ and all $\varepsilon > 0$ there exists an $f \in \mathcal{H}$ with $\| f - g \|_\infty \leq \varepsilon$.*

Consider a binary classification problem, i.e. $\mathcal{Y} = \{-1, +1\}$. The *misclassification risk* of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$\mathcal{R}_{\mathrm{P}}(f) := \mathrm{P}\Big( \{(x,y) \in \mathcal{X} \times \mathcal{Y}; \mathrm{sign} f(x) \neq y\} \Big),$$

where sign(0) := 1. The *Bayes risk* of P is the smallest achievable misclassification risk, *i.e.*

$$\mathcal{R}_{\mathrm{P}} := \inf\left\{\mathcal{R}_{\mathrm{P}}(f); f : \mathcal{X} \to \mathbb{R} \text{ measurable}\right\}.$$

Steinwart (2003, p.1083) proved the following relation between $\mathcal{S}_{L,P}$ and $\mathcal{R}_{\mathrm{P}}$.

**Proposition 9** *Consider a binary classification problem, i.e.* $\mathcal{Y} = \{-1, +1\}$. *Let* $L$ *be a classification calibrated and convex loss function and* P *be a Borel probability measure on* $\mathcal{X} \times \mathcal{Y}$. *Then we have*

$$\mathcal{S}_{L,P} \geq \inf_{f:\mathcal{X}\to\mathbb{R} \text{ measurable}} \left\{\mathrm{P}\left((x,y) \in \mathcal{X}_{cont} \times \mathcal{Y} \text{ with } f(x) \neq y\right)\right\}. \tag{9}$$

*In particular,* $\mathcal{S}_{L,P} \geq \mathcal{R}_{\mathrm{P}}$ *holds if* $\mathcal{X}_{cont} = \mathcal{X}$.

Now we can formulate our result on the number of support vectors.

**Theorem 10 (Number of support vectors)** *Consider an RLB estimator of type I defined by (3). Assume that the assumptions of Theorem 6 are valid. (i) The number of support vectors of the RLB estimator is given by*

$$|\{\alpha_{i,RLB} \neq 0;\, i = 1, \ldots, n\}| = |\{SV(\mathcal{S}_1) \cup \ldots \cup SV(\mathcal{S}_B)\}|\,. \tag{10}$$

*(ii) Consider a binary classification problem, i.e.* $\mathcal{Y} = \{-1, +1\}$. *Let* $B$ *be fixed, and consider* $n := B \cdot n_b \to \infty$. *Let* $L$ *be a classification calibrated and convex loss function,* $k$ *be a universal kernel and* $(\lambda_{n_b})$ *be a sequence of positive regularization parameters with* $\lambda_{n_b} \to 0$ *and* $n_b\lambda_{n_b}^2/|L_{\lambda_{n_b}}|_1^2 \to \infty$, *if* $n_b \to \infty$. *Then for all Borel probability measures* P *on* $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$ *and all* $\varepsilon > 0$ *the RLB-classifier based on (4) with respect to* $k$, $L$, $(\lambda_{n_b})$, *and* $B$ *satisfies*

$$\Pr^{*n}\left(\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_B \in (\mathcal{X} \times \mathcal{Y})^n; \frac{1}{n}|SV(f^{RLB}_{\mathcal{S},B,(\lambda_{n_b})})| \geq S_{L,\mathrm{P}} - \varepsilon\right) \to 1\,. \tag{11}$$

*Here* $\Pr^{*n}$ *denotes the outer probability measure of* $\mathrm{P}^n$ *in order to avoid measurability considerations.*

The result given in (11) has the following interpretation: with probability tending to 1 if the total sample size $n = Bn_b$ converges to $\infty$, but $B$ is fixed, the fraction of support vectors of the kernel based RLB estimator $f^{RLB}_{\mathcal{S},B,(\lambda_{n_b})}(x)$ in a binary pattern recognition problem is essentially greater than the Bayes risk, because (11) is valid for all $\varepsilon > 0$ and $S_{L,P} \geq \mathcal{R}_{\mathrm{P}}$ whenever $\mathcal{X}_{cont} = \mathcal{X}$, see Proposition 9.

Part (ii) of Theorem 10 gives for the RLB estimator the *same* asymptotical bound for the number of support vectors then Steinwart (2003) for $B = 1$ who also gave sharper bounds under different conditions.

Now we investigate conditions to guarantee that RLB estimators using general SVM estimators are $L-$risk consistent, *i.e.* that such *RLB estimators are able to learn*. If P is a probability distribution on $\mathcal{X} \times \mathcal{Y}$, the *L-risk* of a measurable map $f : \mathcal{X} \to \mathbb{R}$ with respect to P is defined by

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int L(Y, f(X))\, d\mathrm{P} = \int L(y, f(x))\, \mathrm{P}(dy|x)\, \mathrm{P}_X(dx).$$

The above integral is always defined since $L$ is non-negative and continuous, although it may be infinite. Consider a general SVM estimator $f_{\mathcal{S},\lambda}$ defined by (4) for the whole data set $\mathcal{S}$. The estimator $f_{\mathcal{S},\lambda_n}$ is called *L-risk consistent*, if

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathcal{S},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}} := \inf\Big\{\mathcal{R}_{L,\mathrm{P}}(f)\,;\; f : X \to \mathbb{R} \text{ measurable}\Big\} \tag{12}$$

holds in probability for $n \to \infty$ for suitable chosen positive regularization sequences $(\lambda_n)_{n\in\mathbb{N}}$. Of course, such convergence can only hold if the used RKHS is rich enough, e.g. if $\mathcal{H}$ is universal.

Several authors have given conditions to guarantee that general SVM estimators are $L-$risk consistent, *cf.* Steinwart (2002, 2005a), Zhang (2004), and Christmann and Steinwart (2005). If $f_{\mathcal{S},\lambda_n}$ is $L$-risk consistent, $B \geq 1$ fixed, and $\lim_{n\to\infty} \min_{1\leq b\leq B} n_b = \infty$, we obtain by Slutzky's theorem for an RLB estimator of type I based on a convex combination with weights $c_b \in (0,1)$ and $\sum_{b=1}^B c_b = 1$ that

$$\sum_{b=1}^{B} c_b\, \mathcal{R}_{L,\mathrm{P}}(f_{\mathcal{S}_b,\lambda_{n_b}}) \to \mathcal{R}^*_{L,\mathrm{P}} \tag{13}$$

in probability for $n \to \infty$. The next result gives $L$-risk consistency of RLB estimators of type I using a convex combination.

**Theorem 11 ($L-$risk consistency)** *Let $f_{\mathcal{S},\lambda_n}$ be an L-risk consistent general SVM estimator based on (4) with a convex loss function and $\lambda_n > 0$. Then the RLB estimator of type I defined by $f^{RLB}_{\mathcal{S},B,(\lambda_{n_b})} = \sum_{b=1}^B c_b f_{\mathcal{S}_b,\lambda_{n_b}}$ with $c_b \in (0,1)$ and $\sum_{b=1}^B c_b = 1$ is L-risk consistent, i.e.*

$$\mathcal{R}_{L,\mathrm{P}}\left(\sum_{b=1}^{B} c_b\, f_{\mathcal{S}_b,\lambda_{n_b}}\right) \quad \xrightarrow{\mathrm{P}} \quad \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty\,. \tag{14}$$

### *3.3 Robustness properties of RLB*

Now we derive results which show that certain robustness properties are inherited from the original estimator $T_{\mathcal{S}}$ to the RLB estimator. We will restrict attention to two robustness approaches. The finite sample breakdown point proposed by Donoho and Huber (1983) measures the worst case behavior of a

statistical estimator. We use the replacement version of this breakdown point, see Hampel *et al.* (1986, p.98). The influence function proposed by Hampel (1968, 1974) measures the impact on the estimation due to an infinitesimal small contamination of the distribution P in direction of a Dirac-distribution.

**Definition 12 (Finite-sample breakdown point)** *Let $\mathcal{S}_n = \{(x_i, y_i), i = 1, \ldots, n\}$ be a data set with values in $\mathcal{X} \times \mathcal{Y}$. The finite-sample breakdown point of an estimator $T_{\mathcal{S}_n}$ is defined by*

$$\varepsilon_n^*(T_{\mathcal{S}_n}) = \max\left\{\frac{m}{n}\,; \text{Bias}(m; T_{\mathcal{S}_n}) \ \text{ is finite}\right\}, \tag{15}$$

*where*

$$\text{Bias}(m; T_{\mathcal{S}_n}) = \sup_{\mathcal{S}'_n} \parallel T_{\mathcal{S}'_n} - T_{\mathcal{S}_n} \parallel \tag{16}$$

*and the supremum is over all possible samples $\mathcal{S}'_n$ that can be obtained by replacing* any *$m$ of the original data points by* arbitrary *values in $\mathcal{X} \times \mathcal{Y}$.*

**Theorem 13 (Finite-sample breakdown point of RLB)** *Consider RLB with $B$ bites where $n_b \equiv n/B$. Denote the finite sample breakdown point of the estimator $T_{\mathcal{S}_b}$ for bite $b$ by $\varepsilon^*_{n_b}(T_{\mathcal{S}_b})$ and denote the finite sample breakdown point of the estimator $\hat{\mu} = \hat{\mu}(T_{\mathcal{S}_1}, \ldots, T_{\mathcal{S}_B})$ in the aggregation step by $\varepsilon^*_B(\hat{\mu})$. Then the finite sample breakdown point of the RLB estimator is given by*

$$\varepsilon^*_{RLB,\mathcal{S},B} = \frac{1}{n}\left(\sum_{b=1}^{k}\left(n_b\varepsilon^*_{n_b}(T_{\mathcal{S}_b}) + 1\right)_{(b:B)} - 1\right), \tag{17}$$

*where $k$ is the smallest integer not less than $B\varepsilon^*_B(\hat{\mu}) + 1$ and $z_{(1:B)} \leq \ldots \leq z_{(B:B)}$ denote the ordered values of $\{z_1, \ldots, z_B\}$.*

**Remark.** If all values $n_b\varepsilon^*_{n_b}(T_{\mathcal{S}_b})$ are equal, we obtain

$$\varepsilon^*_{RLB,\mathcal{S},B} = \frac{\left(n_b\varepsilon^*_{n_b}(T_{\mathcal{S}_b}) + 1\right)\ \lceil B\varepsilon^*_B(\hat{\mu}) + 1\rceil - 1}{n} \geq \varepsilon^*_{n_b}(T_{\mathcal{S}_b})\varepsilon^*_B(\hat{\mu})\,. \tag{18}$$

If the mean or any other estimator with $\varepsilon^*_B(\hat{\mu}) = 0$ is used in this situation, then the RLB has a finite sample breakdown point of $\varepsilon^*_{n_b}(T_{\mathcal{S}_b})/B \to 0$, if $B \to \infty$. Hence $B$ should not be too large. $\lhd$

**Example 14 (Univariate location model)** Consider the univariate location problem, where $x_i \equiv 1$ and $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, $n = 55$. Assume that $y_i \neq y_j$ for $i \neq j$. The finite sample breakdown point of the median is $\lfloor (n-1)/2 \rfloor / n = 27/55 \approx 0.491$. The mean has a finite sample breakdown point of 0. Let us investigate the robustness of RLB with $B = 5$ and $n_b = 11$, $b = 1, \ldots, B$. (a) If the median is used as the location estimator in each bite and if the median is used in the aggregation step, then $\varepsilon^*_{RLB,\mathcal{S}_n,B} = 17/55 \approx 0.309$. This value is reasonably high, but lower than the

finite sample breakdown point of the median for the whole data set. Note that in a *fortunate* situation the impact of up to $(2\times 11+5\times 3)/55 = 0.672$ extreme data points (e.g., all equal to $+\infty$) is still bounded for the RLB estimator in this setup: modify all data points in $B\varepsilon_B^*(\hat{\mu}) = 2$ bites and up to $n_b\, \varepsilon_{n_b}^*(T_{\mathcal{S}_b}) = 5$ data points in the remaining $B(1-\varepsilon_B^*(\hat{\mu})) = 3$ bites. This is no contradiction because the breakdown point measures the *worst case* behavior. (b) If the median is used as the location estimator in each bite and if the *mean* is used in the aggregation step, then we obtain $\varepsilon_{RLB,\mathcal{S}_n,B}^* = \varepsilon_{n_b}^*(T_{\mathcal{S}_b})/B = 5/55 \approx 0.09$. (c) If the mean is used as the location estimator in each bite and also in the aggregation step we have $\varepsilon_{RLB,\mathcal{S}_n,B}^* = 0$. $\lhd$

Now we investigate the influence function of an RLB estimator of type I. We assume the existence of a map $T$ which assigns to every distribution P on a given set $Z$ an element $T(\mathrm{P})$ of a given Banach space $E$ such that our RLB estimator for a data set $S = S_1 \cup \ldots \cup S_B$ has the representation

$$T_{\mathcal{S},B}^{RLB} = \sum_{b=1}^{B} c_b T(\mathrm{P}_{\mathcal{S}_b}). \tag{19}$$

Here $\mathrm{P}_{\mathcal{S}_b}$ denotes the empirical distribution of the data points in bite $\mathcal{S}_b$, $b = 1, \ldots, B$. We have $T(\mathrm{P}) = \theta \in E = \mathbb{R}^d$ for parametric models and $E = \mathcal{H}$ and $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$ for general SVM methods defined by (4).

**Definition 15 (Influence function)** *The influence function of $T$ at a point $z$ for a distribution* P *is the special Gâteaux derivative (if it exists) in direction of the Dirac distribution $\delta_z$, i.e.*

$$\mathrm{IF}(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T\Big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\Big) - T(\mathrm{P})}{\varepsilon}\,. \tag{20}$$

The influence function measures the impact of an infinitesimal small amount of contamination of the probability distribution P in direction of $\delta_z$ on the theoretical quantity of interest $T(\mathrm{P})$. Statistical methods with a *bounded* influence function are desirable. Many robust estimators have a bounded influence function, see e.g. Hampel *et al.* (1986, Chapter 6.3) for $M$-estimators in multivariate location and scale models, and Coakley and Hettmansperger (1993) and Croux *et al.* (2003) for estimators in linear regression models.

**Theorem 16 (Influence function of RLB)** *Assume that the original estimator $T_{\mathcal{S}}$ has the representation $T(\mathrm{P}_n)$, where $\mathrm{P}_n$ is the empirical distribution of the sample $\mathcal{S}$, and that the influence function of the map $T(\mathrm{P})$ exists for the probability distribution* P*. Then the RLB estimator of type I defined by (3) with a fixed number $B$ of bites and $n/n_b \equiv B$ exists and equals the influence function of $T(\mathrm{P})$.*

The influence function of many general SVM methods exists for the case of classification and regression, see Christmann and Steinwart (2004, 2005). Further the influence function of such methods is bounded if a loss function $L$ with bounded first derivative and a bounded and universal kernel $k$ are used. An example is kernel logistic regression in combination with a Gaussian radial basis function kernel $k(x, x') = \exp(-\gamma||x-x'||^2)$, $x, x' \in \mathcal{X}$, where $\gamma \in (0, \infty)$. Some properties of SVM methods for finite sample sizes are given in Christmann *et al.* (2002) and Christmann (2004).

### *3.4 Determination of the number B of bites*

From the results given in Sections 3.1 to 3.3 it is obvious, that the number of bites has some impact on the statistical behavior of the RLB estimator and also on the computation time and the necessary computer memory. An optimal choice of the number $B$ of bites will in general depend on P, but some arguments can be given how to determine $B$ in an appropriate manner.

One should take the sample size $n$, the computer resources (number of CPUs, RAM) and the acceptable computation time into account. The quantity $B$ should be much lower than $n$, because otherwise there is not much hope to obtain useful estimators from the bites and because the finite sample breakdown point is generally decreasing with increasing values of $B$. Further, $B$ should depend on the dimensionality $d$ of the explanatory vectors $x_i \in \mathcal{X}$. *E.g.* a rule of thump for linear regression is that $n/d$ should be at least 5. Hence we propose that $n_b/d \geq 5$ for all bites for linear regression. Because the function $f$ is completely unknown in nonparametric regression assumptions on the complexity of $f$ are crucial. The sample size $n_b$ for each bite should converge to infinity, if $n \to \infty$, to obtain consistency of RLB. The results from some numerical experiments not given here can be summarized as follows. (i) If $B$ is too large, the computational overhead and the danger of bad fits increase because $n_b$ is too small to provide reasonable estimators. (ii) A major decrease in computation time and memory saving is often already present, if $B$ is chosen in a way such that each bite fits nicely into the computer (CPU, RAM). Nowadays robust estimators can often be computed for sample sizes up to $n_b = 10^4$ or $n_b = 10^5$. In this case $B = \lceil n/n_b \rceil$ can be a reasonable choice. (iii) If distribution-free confidence intervals at the $(1-\alpha)$ level for the median of the predictions, i.e. $T^{RLB}_{\mathcal{S},B}(x) = \text{median}_{1 \leq b \leq B} T_{\mathcal{S}_b}(x)$, $x \in \mathcal{X}$, are needed, one should take into account that the actual confidence level of such confidence intervals based on order statistics can be conservative, *i.e.* higher than the specified level, for some pairs $(r, s)$ of order statistics due to the discreteness of order statistics. (iv) In our examples $B = 17$ gave good results.

## 4 Numerical results

In this section we give a few numerical results for RLB. We apply our proposal for a parametric and for a non-parametric method, namely robust linear regression by MM-estimation (Yohai, 1987) and kernel logistic regression (Wahba, 1999). The computations are done on a PC with a 2.8 GHz processor.

Let us begin with robust estimation in linear regression. We simulated data sets with $n = Bn_b$ observations $(x_i, y_i) \in \mathbb{R}^4$ in the following way, where $x_i = (x_{i,1}, x_{i,2}, x_{i,3})$. We consider independent and identically distributed random variables $(X_{i,1}, X_{i,2}, X_{i,3}, Y_i)$, $i = 1, \ldots, n$, each with a probability distribution from a mixture model $\mathrm{P} = 0.8\mathrm{P}_1 + 0.2\mathrm{P}_2$. Under $\mathrm{P}_1$ the explanatory variables $X_{i,1}$, $X_{i,2}$, and $X_{i,3}$ are independent and have a Student $t$-distribution with 3 degrees of freedom and location parameter 0, whereas the response variable $Y_i|(x_{i,1}, x_{i,2}, x_{i,3})$ has a Student $t$-distribution with 3 degrees of freedom and location parameter $f(x_i) = x_i^T\theta$, where $x_i^T = (0, x_{i,1}, x_{i,2}, x_{i,3})$ and $\theta = (0, 1, 1, 1)^T$. Hence $\mathbb{E}_{\mathrm{P}_1} X_{i,j} = 0$ and $\mathbb{E}_{\mathrm{P}_1}[Y_i|(x_{i,1}, x_{i,2}, x_{i,3})] = f(x_i)$, $1 \leq i \leq n$. Under $\mathrm{P}_2$ the explanatory variables $X_{i,1}$, $X_{i,2}$, and $X_{i,3}$ are independent and identically distributed each with a Dirac distribution in the point 50, whereas the response variable $Y_i|(x_{i,1}, x_{i,2}, x_{i,3})$ has a Dirac distribution in the point 1000, i.e. $\mathrm{P}_2((X_{i,1}, X_{i,2}, X_{i,3}, Y_i) = (50, 50, 50, 1000)) = 1$. Obviously the distribution P produces approximately 20% bad leverage points in $(50, 50, 50, 1000)$ with respect to the linear regression model $\mathrm{P}_1$. Table 2 shows the computation times in seconds, the bias of an MM-estimator computed for the whole data set and of the RLB estimator for $B = 17$, and the width of the componentwise distribution-free confidence intervals based on the median at the 95%-level for different sub-sample sizes $n_b$. The MM-estimates were computed with the function `rlm` from the R-library `MASS` (Venables and Ripley, 2002). This function first computes an S-estimate as a starting point which has an approximate finite sample breakdown point of 0.5. Then an M-estimator with Tukey's biweight and fixed scale is iteratively computed using this starting value that will inherit this breakdown point from the S-estimator. The time-consuming phase of MM-estimators is the computation of the highly robust starting value. The confidence intervals for the original MM-estimator were computed using the asymptotical normality assumption. The distribution-free confidence intervals for the RLB estimator were based on the $5^{th}$ and the $13^{th}$ order statistics. Because the bias terms and the width of the confidence intervals are very small due to the large sample size, the values in Table 2 are multiplied by $10^3$.

In the considered situations RLB gave good results: the bias values are small, which shows that the RLB method indeed gave robust estimates, and the width of the confidence intervals is of similar size as for the original MM-estimator. It is not surprising that the distribution-free confidence intervals

for the RLB estimator are somewhat larger (often by a factor between 1.1 and 1.2) than the confidence intervals of the MM-estimator based on the assumption of asymptotic normality. If the total sample size $n$ is not too big, such that the MM-estimates can be computed with `rlm` only using the RAM of the computer, RLB only saves a little bit of computation time. However, using RLB one can fit for much larger data sets for which the algorithm used by `rlm` would need much more RAM than the available PC has (2 GB), such that the computation of the MM-estimates for the whole data set was impossible. In contrast to that, the computation time of RLB increased only approximately linearly in $n_b$, and the used RAM was low in contrast to the used RAM to compute the MM-estimates for the whole data set. No memory problems occurred for RLB with $n = 3.4$ million and $B = 17$.

| | $n_b = 10,000$ | | $n_b = 100,000$ | | $n_b = 200,000$ | |
|---|---|---|---|---|---|---|
| | RLB | MM | RLB | MM | RLB | MM |
| seconds | 33.89 | 44.64 | 348.78 | 460.95 | 684.61 | – |
| Bias($\hat{\theta}_0$) ($\times$1000) | 2.32 | 0.35 | 0.17 | 0.17 | 0.31 | – |
| width of c.i. ($\times$1000) | 17.42 | 15.36 | 5.15 | 4.87 | 5.27 | – |
| Bias($\hat{\theta}_1$) ($\times$1000) | 1.21 | 1.18 | −2.02 | −1.44 | 0.46 | – |
| width of c.i. ($\times$1000) | 8.78 | 7.39 | 3.29 | 2.31 | 1.39 | – |
| Bias($\hat{\theta}_2$) ($\times$1000) | 0.62 | 0.23 | 0.09 | −0.32 | 0.90 | – |
| width of c.i. ($\times$1000) | 8.06 | 7.38 | 2.32 | 2.30 | 2.82 | – |
| Bias($\hat{\theta}_3$) ($\times$1000) | −1.60 | −2.22 | 0.31 | −0.16 | −0.54 | – |
| width of c.i. ($\times$1000) | 8.72 | 7.36 | 5.19 | 2.28 | 1.86 | – |

Table 2
Results for robust linear regression with MM-estimator and RLB with $B = 17$. The computation of the MM-estimates for the whole data set with $n = 17 \cdot 200,000 = 3.4$ million data points was not possible due to memory problems.

Now we apply the RLB approach to kernel logistic regression (KLR), see (Wahba, 1999). KLR is a flexible method for classification problems and provides also estimates for the conditional probabilities $\mathrm{P}(Y = 1|X = x)$, $x \in \mathcal{X}$, which is not true for the support vector machine, see Bartlett and Tewari (2004). Christmann and Steinwart (2004) showed KLR has good robustness properties, e.g. a bounded influence function. All computations are done with the program `myKLR` (Rüping, 2003) which is an implementation in C++ of the algorithm proposed by Keerthi *et al.* (2004) to solve the dual problem. We choose KLR for two reasons. Firstly, the computation of KLR needs much more time than for the support vector machine because the latter solves a quadratic instead of a convex program in dual space. Therefore, the need for computational improvements is greater for KLR than say for the SVM, and the potential gains of RLB can be more important. Secondly, the number of support vectors of KLR is approximately equal to $n$ which slows down the computation of predictions.

The simulated data sets contain $n$ data points $(x_i, y_i) \in \mathbb{R}^8 \times \{-1, +1\}$ simulated in the following way. All 8 components of $x_i = (x_{i,1}, \ldots, x_{i,8})$ are simulated independently from a uniform distribution on $(0, 1)$. The responses $y_i$ are simulated independently from a logistic regression model according to $\mathrm{P}(Y_i = +1 | X_i = x_i) = 1/(1 + \exp[-f(x_i)])$. We define

$$f(x_i) = \sum_{j=1}^{8} x_{i,j} - x_{i,1}x_{i,2} - x_{i,2}x_{i,3} - x_{i,4}x_{i,5} - x_{i,1}x_{i,6}x_{i,7} \,,$$

such that there are 8 main effects and 4 interaction terms. The numerical results of fitting kernel logistic regression to such data sets is given in Table 3. If the whole data set has $n = 10^5$ observations, approximately 10 hours were needed to compute KLR. If RLB with $B = 10$ bits is used each with a sub-sample size of $n_b = 10^4$, one needs approximately $10 \times 93.3$ seconds, *i.e.* 16 minutes, if 1 GB of kernel cache is available. This is a reduction by a factor of 38. If there are 5 CPUs available and each processor can use up to 200 MB kernel cache, RLB with $B = 10$ will need approximately 11 minutes which is a reduction by a factor of 55.

| sample size $n$ | CPU time | used cache in MB | available cache in MB |
|---:|---:|---:|---:|
| 2,000 | 4 sec | 33 | 200 |
| 5,000 | 25 sec | 198 | 200 |
| 10,000 | 5 min, 21 sec | 200 | 200 |
| 10,000 | 1 min, 33 sec | 787 | 1,000 |
| 20,000 | 24 min, 11 sec | 1,000 | 1,000 |
| 100,000 | 9 h, 56 min, 46 sec | 1,000 | 1,000 |

Table 3
Computation times for kernel logistic regression using `myKLR`.

Christmann (2005) described a strategy to construct insurance tariffs for a data set from 15 German motor vehicle insurance companies. The whole data contains data from around 4.6 million customers with dozens of explanatory variables. A direct use of kernel logistic regression with `myKLR` was unfeasable due to the computation time, see Table 3. Although a strategy was used to reduce the computational effort by exploiting characteristic features of such data sets, RLB can substantially reduce the computation time. Fitting the model to the whole data set would need more than six months on a PC with 2.8 GHz CPU, whereas RLB with $B = 17$ using 2 CPUs was able to provide a good fit within 4 days, which is a reduction by a factor of 45. If RLB is allowed to use 8 CPUs one expects that the computation can be done in approximately one day. This turned out to be true when we used RLB on a LINUX cluster. RLB is therefore quite useful for kernel logistic regression for large data sets.

## 5 Discussion and related work

In this paper robust learning from bites (RLB) was proposed to broaden the usability of computer-intensive robust methods in the case of large data sets which occur nowadays often in data mining projects. RLB is especially designed for situations under which the original robust method cannot be used due to excessive computation time or due to memory space problems. In these situations RLB offers robust estimates and additionally robust confidence intervals. RLB estimators will in general not fulfill certain optimality criteria, but the method has four nice properties. *Scalability:* the number $B$ of bites can be chosen such that the algorithm used to fit the bites needs less memory than the computer offers. *Performance:* the computational steps for different bites can easily be distributed on several processors because they are independent and use disjoint parts of the data set. *Robustness:* we considered the finite sample breakdown point and the influence function. These properties are inherited from the original robust estimator computed for each bite and from the location estimator used to aggregate the results from the bites. *Confidence intervals:* no complex formulae are needed to obtain distribution-free (componentwise) confidence intervals for the estimates or for the predictions if the median is used in the aggregation step because the estimators computed from the $B$ bites are independent and identically distributed. Such confidence intervals for the predictions are especially interesting for general SVM methods (*e.g.* support vector machines and kernel logistic regression), because such methods have nice properties but finite sample confidence intervals for the predictions based on applying such methods once for the whole data set are typically unknown.

Some good robust estimators are not $n^{-1/2}$-consistent and have a complicated non-normal limiting distribution, see e.g. Rousseeuw (1984), Davies (1990), Kim and Pollard (1990), Rousseeuw and Hubert (1999), Bai and He (1999), Van Aelst *et al.* (2002), and Zuo and Cui (2005). Then RLB can be useful if distribution-free confidence intervals for the median of the predictions are needed for large data sets.

Recently Croux *et al.* (2003) proposed estimators with a bounded influence function in a linear regression model to obtain robust standard errors and robust estimators for the covariance matrix of the regression parameters. We like to mention that it is possible to use RLB also for these purposes, if the original estimator has the desired properties, say asymptotic normality and a positive breakdown point or a bounded influence function, and if the data set is large. This follows for a large class of RLB estimators of type I from Proposition 4, Theorem 13, and Theorem 16.

We like to mention that it is possible to use RLB also for these purposes, if

the original estimator has the desired robustness properties and if the data set is large. E.g. consider a consistent and robust estimator for the covariance matrix. Then the corresponding RLB estimator inherits consistency and robustness from the original estimator, see Propositions 4 and 5 and Theorems 13 and 16, but the computation of RLB will be faster than for the original method for large data sets due to Proposition 2.

The subsampling approach used by RLB has connections to several other methods. One example is the remedian proposed by Rousseeuw and Bassett (1990) for univariate location estimation. The remedian with base $B$ computes medians of groups of $n_b$ observations, and then the medians of these medians etc., until only a single estimate remains. The remedian needs only $O(\log(n))$ total storage for fixed $B$ which makes it especially useful for robust estimation in large data bases, for real-time engineering applications in which the data are not present at the same time and perhaps not stored, and for resistant aggregation of curves. As one referee pointed out Balakrishnan and Madigan (2006, p.20f) use a sequential method called streaming to build sparse generalized linear models for large data sets, but the application of this approach to kernel based methods seems to be open research. It is not yet known whether the streaming method offers the same robustness properties than RLB.

RLB has also similarities to Rvote proposed by Breiman (1999), DRvote with classification trees using majority voting proposed by Chawla *et al.* (2004), and with stochastic gradient boosting and greedy function approximation (Friedman, 2001, 2002) which are implemented in the software TreeNet. TreeNet is often able to offer a model with high predictive power. The interpretation of the results can be difficult because a model determined by TreeNet consists typically of more than hundred small trees each with two to eight terminal nodes. As one referee pointed out the computation of trees is fast, trees seem to have good robustness properties, and the combination of MART and random forests is often successful in data mining. However, mathematical proofs that these methods share the same robustness properties and the $L-$risk consistency with RLB are not yet available to our knowledge.

There exist also relationships between RLB, cross validation, and robust bootstrapping methods described e.g. by Amado and Pires (2004), Salibian-Barrera *et al.* (2005), and Willems and Van Aelst (2005). However, cross validation and robust bootstrapping of computer-intensive methods for huge data sets is not always a simple task due to computation time and memory limitations of the computer. That was one motivation for the present work.

The focus of the present paper was on robustness aspects and the computation of robust distribution-free confidence intervals for the median of the predictions even for very large data sets. Such confidence intervals are often a problem for robust estimators and general SVM methods based on Vap-

nik's convex risk minimization principle. These topics were not covered in the papers mentioned above. RLB has also some similarity to the algorithms FAST-LTS and FAST-MCD developed by Rousseeuw and Van Driessen (1999, 2000) for robust estimation in linear regression or multivariate location and scatter models for large data sets. FAST-LTS and FAST-MCD split the data set into sub-samples, optimize the objective function in each sub-sample, and use these solutions as starting values to optimize the objective function for the whole data set. This is in contrast to RLB which aggregates estimation results from the bites to obtain robust confidence intervals.

## A Appendix

The appendix contains the proofs for the results given in Section 3.

**Proof of Proposition 2**. Obvious. □

**Proof of Proposition 3**. Obvious. □

**Proof of Proposition 4**. (i) follows from the linearity of the expectation operator. (ii) and (iii) follow from Slutzky's theorem. □

**Proof of Proposition 5**. By construction of RLB the bites are disjoint and the estimators from the bites are independent. Assume that the original estimator $T_{\mathcal{S}}$ is consistent in probability for $T_{\mathrm{P}}$. We have for all $\varepsilon > 0$ that

$$\begin{aligned}
&\mathrm{P}\left(|\mathrm{median}_{b=1,\ldots,B}\, T_{\mathcal{S}_b}(x) - T_{\mathrm{P}}(x)| < \varepsilon\right) \\
\geq\, &\mathrm{P}\left(|T_{\mathcal{S}_b}(x) - T_{\mathrm{P}}(x)| < \varepsilon \text{ for all } b = 1, \ldots, B\right) \\
=\, &\prod_{b=1}^{B} \mathrm{P}\left(|T_{\mathcal{S}_b}(x) - T_{\mathrm{P}}(x)| < \varepsilon\right) \to 1, \quad n \to \infty\,, \quad x \in \mathcal{X}\,,
\end{aligned}$$

because $B$ is fixed and $\lim_{n\to\infty}\,(n/n_b) = B$. Now, assume that the original estimator $T_{\mathcal{S}}$ is strongly consistent to $T_{\mathrm{P}}$. Then we obtain analogously:

$$\mathrm{P}\left(\lim_{n\to\infty} \mathrm{median}_{b=1,\ldots,B}\, T_{\mathcal{S}_b}(x) = T_{\mathrm{P}}(x)\right) \geq \prod_{b=1}^{B} \mathrm{P}\left(\lim_{n_b\to\infty} T_{\mathcal{S}_b}(x) = T_{\mathrm{P}}(x)\right) = 1. \;\square$$

**Proof of Theorem 6**. By assumption each bite $\mathcal{S}_b$ is fitted with a general

SVM estimator having the representation

$$f_{\mathcal{S}_b,\lambda}(x) = \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i), \quad i \in \mathcal{S}_b\,,\ b = 1, \ldots, B,\ x_i \in \mathcal{X}.$$

Because the bites $\mathcal{S}_b$, $b = 1, \ldots, B$, are disjoint, the RLB estimator using the mean in the aggregation step is given by

$$f^{RLB}_{\mathcal{S},B,\lambda}(x) = \sum_{b=1}^{B} c_b \sum_{i=1}^{n_b} \alpha_{i,b}\, k(x, x_i) = \sum_{i=1}^{n} \sum_{b=1}^{B} c_b\, \alpha_{i,b}\, k(x, x_i),\ x \in \mathcal{X}.$$

The formula (8) follows immediately. □

**Proof of Theorem 10**. (i) follows from (8). (ii) Steinwart (2003, Th.9) proved

$$\Pr^{*n}\left(\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^n; |\mathrm{SV}(f_{\mathcal{S},\lambda_n})| \geq (S_{L,\mathrm{P}} - \varepsilon)n\right) \to 1\,, \quad n \to \infty\,. \tag{A.1}$$

Denote the outer probability measure of the product measure $\mathrm{P}^{b,n_b}$ by $\Pr^{*b,n_b}$. The bites $\mathcal{S}_b$, $b = 1, \ldots, B$, are independent and identically distributed by construction of RLB. Using (A.1) and $n \equiv Bn_b$ we obtain

$$\begin{aligned}
&\Pr^{*n}\left(\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^n; |\mathrm{SV}(f^{RLB}_{\mathcal{S},B,(\lambda_{n_b})})| \geq (S_{L,\mathrm{P}} - \varepsilon)n\right) \\
&= \Pr^{*n}\left(\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_B \in (\mathcal{X} \times \mathcal{Y})^n; |\mathrm{SV}(f_{\mathcal{S}_b,\lambda_{n_b}})| \geq \sum_{b=1}^{B} (S_{L,\mathrm{P}} - \varepsilon)n_b\right) \\
&\geq \Pr^{*n}\left(\forall\ \mathcal{S}_b \in (\mathcal{X} \times \mathcal{Y})^{n_b}, b = 1, \ldots, B\,;\ |\mathrm{SV}(f_{\mathcal{S}_b,\lambda_{n_b}})| \geq (S_{L,\mathrm{P}} - \varepsilon)n_b\right) \\
&= \prod_{b=1}^{B} \Pr^{*b,n_b}\left(\mathcal{S}_b \in (\mathcal{X} \times \mathcal{Y})^{n_b}; |\mathrm{SV}(f_{\mathcal{S}_b,\lambda_{n_b}})| \geq (S_{L,\mathrm{P}} - \varepsilon)n_b\right) \to 1\,,\ n \to \infty.\ \square
\end{aligned}$$

**Proof of Theorem 11**. Note that $f^{RLB}_{\mathcal{S},B,(\lambda_{n_b})}$ of type I is a convex combination of $f_{\mathcal{S}_b,\lambda_{n_b}}$, $b = 1, \ldots, B$, due to (3). We obtain for $\lim_{n\to\infty} \min_{1\leq b\leq B} n_b = \infty$:

$$\begin{aligned}
0 &\leq \int L\left(Y, \sum_{b=1}^{B} c_b f_{\mathcal{S}_b,\lambda_{n_b}}(X)\right)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \int \sum_{b=1}^{B} c_b L\left(Y, f_{\mathcal{S}_b,\lambda_{n_b}}(X)\right)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}} \qquad (A.2) \\
&= \sum_{b=1}^{B} c_b \left[\int L\left(Y, f_{\mathcal{S}_b,\lambda_{n_b}}(X)\right)\, d\mathrm{P} - \mathcal{R}^*_{L,\mathrm{P}}\right] \quad \xrightarrow{\mathrm{P}}\ 0, \qquad (A.3)
\end{aligned}$$

Here we used the convexity of $L$ in (A.2) and the $L$-risk consistency of the original estimator in (A.3). □

**Proof of Theorem 13**. The minimum number of points needed to modify $T_{\mathcal{S}_b}$ in bite $b$ such that the bias in (16) is infinite is given by $n_b \cdot \varepsilon^*_{n_b}(T_{\mathcal{S}_b}) + 1$, $b = 1, \ldots, B$. The RLB estimator breaks down if at least $B\varepsilon^*_B(\hat{\mu}) + 1$ of the estimators $T_{\mathcal{S}_1}$, ..., $T_{\mathcal{S}_B}$ break down. Hence, we can modify $(n_b \cdot \varepsilon^*_{n_b}(T_{\mathcal{S}_b}) + 1)(B\varepsilon^*_B(\hat{\mu}) + 1) - 1$ data points in an arbitrary way without obtaining an infinite bias. This gives the assertion. □

**Proof of Theorem 16**. Let $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and P be a probability distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$. By assumption the RLB estimator fulfills (19) and the influence function $\mathrm{IF}(z; T, \mathrm{P})$ exists. It follows

$$\begin{aligned}
\mathrm{IF}(z; T_B^{RLB}, \mathrm{P}) &= \lim_{\varepsilon \downarrow 0} \frac{T_B^{RLB}\Big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\Big) - T_B^{RLB}(\mathrm{P})}{\varepsilon} \\
&= \lim_{\varepsilon \downarrow 0} \frac{\sum_{b=1}^B c_b T\Big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\Big) - \sum_{b=1}^B c_b T(\mathrm{P})}{\varepsilon} \\
&= \sum_{b=1}^B c_b \lim_{\varepsilon \downarrow 0} \frac{T\Big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\Big) - T(\mathrm{P})}{\varepsilon} = \mathrm{IF}(z; T, \mathrm{P}). \qquad \square
\end{aligned}$$

## References


Amado, C. and Pires, A. (2004). Robust bootstrap with non random weights based on the influence function. *Communications in Statistics, Simulation and Computation*, **33**(2), 377–396.

Bai, Z. and He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Annals of Statistics*, **27**, 1616–1637.

Balakrishnan, S. and Madigan, D. (2006). Algorithms for sparse linear classifiers in the massive data setting. University of Rutgers, preprint, `http://www.stat.rutgers.edu/~madigan/PAPERS/sm.pdf`.

Bartlett, P. and Tewari, A. (2004). Sparseness versus estimating conditional probabilities: Some asymptotic results. In *Proceedings of the 17th Annual Conference on Learning Theory, Vol. 3120*, Lecture Notes in Computer Science, Vol. 3120, pages 564–578, New York. Springer.

Bartlett, P., Jordan, M., and McAuliffe, J. (2006). Convexity, classification, and risk bounds. *To appear in: Journal of Machine Learning Research*. `http://stat-www.berkeley.edu/tech-reports/638.pdf`.

Breiman, L. (1999). Pasting bites together for prediction in large data sets. *Machine Learning*, **36**, 85–103.

Chawla, N., Hall, L., Bowyer, K., and Kegelmeyer, W. (2004). Learning ensembles for bites: a scalable and accurate approach. *Journal of Machine Learning Reseach*, **5**, 421–451.

Christmann, A. (2004). On properties of support vector machines for pattern recognition in finite samples. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, Series: Statistics for Industry and Technology, pages 49–58. Birkhäuser.

Christmann, A. (2005). On a strategy to develop robust and simple tariffs from motor vehicle insurance data. *Acta Mathematicae Applicatae Sinica (English Series, Springer)*, **21**, 193–208.

Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, **5**, 1007–1034.

Christmann, A. and Steinwart, I. (2005). Consistency and robustness of kernel based regression. University of Dortmund, SFB-475, TR-01/05. Submitted.

Christmann, A., Fischer, P., and Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, **17**, 273–287.

Coakley, C. and Hettmansperger, T. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, **88**, 872–880.

Croux, C., Van Aelst, S., and Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *Annals of the Institute of Statistical Mathematics*, **55**, 265–285.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, 3rd ed.* Wiley & Sons, New York.

Davies, P. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, **18**, 1651–1675.

Donoho, D. and Huber, P. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, and J. Hodges, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Jr., Belmont, California, Wadsworth.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367–378.

Hampel, F. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Dept. Statistics, Univ. California, Berkeley.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*.

MIT Press, Cambridge, Massachusetts.

Hipp, J., Güntzer, U., and Grimmer, U. (2001). Data quality mining - making a virtue of necessity. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD, Santa Barbara, CA, `http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf`.

Keerthi, S., Duan, K., Shevade, S., and Poo, A. (2004). A fast dual algorithm for kernel logistic regression. In *Machine Learning: Proceedings of the Ninetheenth International Conference (Eds: C. Sammut, A.G. Hoffmann)*, pages 299–306. Kaufmann, San Francisco.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, **18**, 191–219.

Le Cam, L. (1980). In: Discussion of "minimum chi-square, not maximum likelihood! *Annals of Statistics*, **8**, 473–478.

Rousseeuw, P. and Bassett, G. (1990). The remedian: a robust averaging method for large data sets. *Journal of the American Statistical Association*, **85**(409), 97–104.

Rousseeuw, P. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, **94**, 388–402.

Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Rousseeuw, P. and Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In W. Gaul, O. Opitz, and M. Schader, editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346, New York. Springer.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Rüping, S. (2003). myKLR - kernel logistic regression. University of Dortmund, Department of Computer Science, `http://www-ai.cs.uni-dortmund.de/SOFTWARE`.

Salibian-Barrera, M., Van Aelst, S., and Willems, G. (2005). PCA based on multivariate MM-estimators with fast and robust bootstrap. *To appear in: Journal of the American Statistical Association*.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.

Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, **18**, 768–791.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, **4**, 1071–1105.

Steinwart, I. (2005a). Consistency of support vector machines and other regularized kernel machines. *IEEE Transactions on Information Theory*, **51**, 128–142.

Steinwart, I. (2005b). How to compare loss functions and their risks. Technical

Report LA-UR-05-7016, Los Alamos National Laboratory. http://www.c3.lanl.gov/~ingo/pubs.shtml.

Steinwart, I., Hush, D., and Scovel, C. (2006). Function classes that approximate the bayes risk. In *COLT'06, 19th Conference on Learning Theory*, Pittsburgh.

Van Aelst, S., Rousseeuw, P., M.Hubert, and Struyf, A. (2002). The deepest regression method. *Journal of Multivariate Analysis*, **81**, 138–166.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S,* $4^{th}$*ed.*. Springer.

Wahba, G. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88, Cambridge, MA. MIT Press.

Willems, G. and Van Aelst, S. (2005). Fast and robust bootstrap for LTS. *Computational Statistics and Data Analysis*, **48**(4), 703–715.

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, **15**, 642–656.

Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56–134.

Zuo, Y. and Cui, H. (2005). Depth weighted scatter estimators. *Annals of Statistics*, **33**, 381–413.